图书情报工作
第 62 卷 第 1 期　2018 年 1 月

基于用户自然标注的 TF-IDF 辅助标引算法及实证研究*

■ 陈白雪　宋培彦
中国科学技术信息研究所　北京 100038

摘要：［目的/意义］从用户角度出发，研究基于用户自然标注的 TF-IDF 辅助标引算法。［方法/过程］首先以核心期刊论文中作者标注的关键词和分类号为源数据，通过对关键词词频进行统计，使用 TF-IDF 算法构建用户标注词表、形成标引知识库，然后通过 IK Analyzer 分词软件对待标引的科技项目数据进行切词和停用词处理，进而使用 TF-IDF 算法和位置加权算法提取科技项目数据的特征词，最终实现对科技项目数据进行关键词和分类的同步标引。［结果/结论］实验结果表明，机标关键词与人标关键词的相似比在 60% 以上的科技项目数据占总数的 68.1%，机标分类号与人标分类号前三位一致的占总数的 83.9%，结果表明基于用户自然标注数据并采用 TF-IDF 算法在关键词和分类标引方面是可行的。

关键词：辅助标引　用户自然标注　TF-IDF 算法　信息组织
分类号：G254
DOI：10.13266/j.issn.0252－3116.2018.01.017

1　引言

信息组织是根据信息资源检索的需要，以文本及各类型的信息源为对象，通过对其内容特征等的分析、选择、标引、处理，使其成为有序化集合的活动[1]。其中，信息标引就是对信息内容进行分析并充分而有效地予以揭示。信息标引分主题标引和分类标引，主题标引是依据特定的主题，赋予文献主题标识的过程，主题标引可以采用标题语言、叙词语言和关键词语言等；分类标引是依据特定的分类语言，赋予文献分类标识的过程。大数据环境下，机器往往需要依据相关的知识库，从文本中抽取能够表达文献信息内容的关键词或分类号，用于文本检索和分类导航等方面，因此，知识库的构建是自动标引的重要研究内容之一。

通过用户自然标注构建知识库是自动标引的一个重要思路。用户自然标注是用户在无意中为自然语言处理研究的各种资源作了一定程度的义务"标注"，是因特网用户对自己的资源或收藏的他人资源添加标签的活动，标签是用户选取的、代表被标注资源的符号，可以是文字，也可以是其他符号[2-3]。用户自然标注是根据用户已有的知识，结合对文献内容的理解，给出能够代表该文本主要内容的标签或词语。目前，自动标引抽出的表达文献主题的关键词的准确性偏低，这在一定程度上是因为自动标引使用的知识库通常是依靠领域专家手工建立的，难以较为全面地将用户使用的词语包含进去，其覆盖面和更新速度有待提高。而用户自然标注能够为扩充知识库提供一个途径，将用户对某一领域内常用的概念或主题词全面快速地扩充，并尽可能符合用户的使用习惯。因此，研究基于用户自然标注的机器辅助标引算法，在提高自动标引的准确率以及标引结果更加符合用户使用习惯方面具有重要意义。

2　相关研究

2.1　标引相关研究

国内外对自动标引的研究主要集中在标引算法的研究。章成志[4]整合了统计机器学习模型与集成学习方法的优势，并结合多分类模型投票的方式，对文档进

* 本文系 2016 年国家社会科学基金项目"基于知识组织的科研项目评审专家发现研究"（项目编号：16BTQ079）和 2017 年度中国科学技术信息研究所创新研究基金面上项目"面向国家科技大数据的知识图谱动态构建方法研究"（项目编号：MS2017－06）研究成果之一。
作者简介：陈白雪（ORCID：0000－0003－4726－8103），研究实习员，硕士，E－mail：chenbx@istic.ac.cn；宋培彦（ORCID：0000－0003－1055－2717），副研究馆员，博士，硕士生导师。
收稿日期：2017－07－10　修回日期：2017－10－23　本文起止页码：132－139　本文责任编辑：王善军

行自动标引;李纲[5]等利用基于知网的词语语义相关算法对词汇链的构建算法进行了改进,并结合词频和词的位置等统计信息,进行关键词的自动标引;曹树金等[6]以逸仙时空BBS为舆情信息源,设计了主题帖自动标引和情感倾向性分析策略,并对主题帖自动标引结果、倾向性人工判断与自动分析的结果进行了对比;王丹等[7]针对中文自动标引过程中出现的歧义词现象,提出一种将穷举法和消歧规则相结合的歧义词消除方法,并验证了该方法的有效性;L. M. De Campos等[8]运用贝叶斯网络对叙词表进行建模,并使用概率推理,选择出最能描述待分类文档的描述符集合,对待分类文档进行自动标引和分类;O. Medelyan等[9]通过从特定领域叙词表中收集术语和短语的语义信息来提高关键词的自动抽取;Z. A. Merrouni等[10]对现有的关键词自动抽取方法进行了概述,并分析各种方法的优点和不足。通过以上研究,学者多采用统计方法、语义算法以及机器学习等方法对文本信息进行自动标引,在标引过程中,对词表或分类表的要求较高,制约了适用范围。基于用户自然标注,通过构建用户自然标注词表,进而优化标引算法,有望提高标引效率和质量。

2.2 用户标注相关研究

国内外对用户标注的研究主要集中在用户标注语义模型、用户标注行为等方面。在用户语义标注模型方面,白化[2]通过建立用户标注模型和语义联系,使用元数据与本体语言对用户标注进行语义描述,使之成为标签本体,适应新一代网络的发展。在用户标注行为研究方面,李枫林等[11]通过对用户标注行为分析,详细研究了用户标注行为所反映的网页间相关性、标签间相关性以及网页和标签间相关性的关联程度,并将这种相关性用于标签相关性计算上,改进了SPR算法;吴丹等[12]以武汉大学图书馆和豆瓣网为例,通过真实的用户日志数据比较二者的用户标注行为,为图书馆更好地开展图书标注服务提出建议;谢佳琳等[13]基于图书馆标注系统质量的视角,以信息系统成功模型为框架构建模型,研究了信息质量、系统质量、服务质量、后悔以及满意对高校图书馆用户标注行为的影响;J. Patterso等[14]通过特定的方式对用户标注内容进行显示和隐藏,设计出了适合学生使用的电子书系统,为用户推荐合适的电子书;M. A. Zarro[15]创建了个人和历史记忆、其他资源链接、修改、翻译等4种用户标注类型来了解图书馆用户的意图,以及在搜索、内容描述和信息检索的方面的影响;Y. Zhang等[16]分析了学术博客中标签的内容特征,根据标签的内容和使用频率,分析学术用户的专业类型;X. Pan等[17]通过构建超网络,分析不同流行标签的使用模式、规律与用户活动的关系及社会标签资源的兴趣水平,用于发现小群志同道合的用户,同时在识别有趣资源方面能发挥积极作用。

另外,马费成等[18]利用标签分析和确定概念的序化和聚类,揭示了用户在图书标注环境下的认知特征;常唯[19]对对网络环境下的用户标注进行了探析,讨论了用户标注在资源组织、异构资源整合、协同过滤和推荐等方面的应用,进而分析其在资源创建、揭示资源内容、记录隐形知识、评价资源等方面的应用价值。

以上学者从用户标注模型、用户标注行为以及用户标注聚类等角度对用户标注相关的内容进行了研究,并取得了一定的成果。用户标注能够反映用户的意图,体现用户对特定知识领域的认知和使用习惯,有利于将用户标注的内容应用在知识组织、异构资源整合、信息推送等方面。然而,传统的用户标签通常依靠手工标引,标引效率偏低。通过用户自然标注数据,能够研究不同用户对某一特定主题的描述方式、表达习惯,从而对用户标注的内容进行分类和聚类,从用户角度对信息内容进行组织和分类,实现对不同来源资源的整合,提高用户的标引效率和检索效率。

2.3 TF-IDF算法相关研究

TF-IDF(term frequency-inverse document frequency)是一种用于信息检索与数据挖掘的常用加权技术[20]。国内外对TF-IDF算法的研究主要集中在算法改进上,路永和[21]等通过将权重修正函数(TW)与TF-IDF结合作为新的特征权重算法,用于文本分类;覃世安[22]等利用特征值在类间出现的概率比代替特征值在类间出现的次数比改进TF-IDF算法,并配合简单累加求和的分类器,用于提高网页文本分类的准确率;刘勘等[23]根据特征词的词频、所在位置和词性提出了改进TF-IDF特征词加权算法的科技文献聚类方法;A. B. Samoylov[24]通过将基于规则的方法和标准词袋模型相结合的方法,用于评估语义分析中ΔTF-IDF特征值;S. Philip[25]将TF-IDF与余弦相似性度量相结合,提出一种基于用户查询的推荐算法;R. Xu[26]针对词性对检索结果的影响,提出了基于词性加权的TF-IDF算法,并将该算法应用在MOOC的搜索引擎中,取得了非常积极的结果;S. M. H. Dadgar等[27]提出了TF-IDF与SVM相结合的文本分类方法,用于社交网站中的新闻分类,并验证了该方法的有效性。

通过对TF-IDF算法的相关研究可以发现,TF-IDF

算法在文本分类方面应用较为广泛,操作简单,易于改进,是提取文本特征常用算法之一。因此,笔者试图将用户标签作为知识来源,采用 TF-IDF 算法构建知识库,并将该算法与位置加权算法相结合,用于提取文本内容的特征词,通过知识库支撑信息标引和文本分类,实现用户标签与信息标引的有效结合。

3　辅助标引算法研究框架

3.1　整体框架设计

为对待标引数据进行关键词和分类号的同步标引,笔者选择了中文核心期刊中的科技论文作为“用户自然标注词表”的数据来源。在科技论文中,作者为每篇论文赋予了关键词和分类号,在多数情况下,这些关键词和分类号是由作者自由标注的。另外,作者作为科研共同体,既是用户标注数据的生产者,又是科研数据的使用者,尤其是核心期刊的论文作者,具有较高的学术素养,专业性较强,数据标注质量较高;同时,论文采用的分类法多为国内通用的《中国图书馆分类法》,规范性较强,因此,用核心期刊论文的关键词和分类号来构建用户自然标注词表是可行的。基于用户自然标注的 TF-IDF 辅助标引算法的技术路线如图 1 所示:

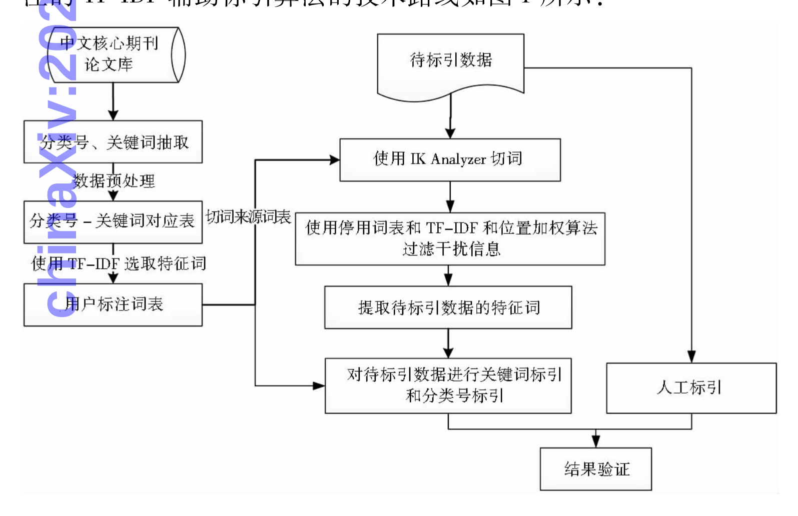


图 1　基于用户自然标注的 TF-IDF 辅助标引算法技术路线图

基于用户自然标注的 TF-IDF 辅助标引算法,以中文核心期刊论文库为语料库,抽取论文的关键词和分类号,通过对关键词和分类号使用 TF-IDF 算法,构建用户标注词表;以用户标注词表为基础,对待标引数据进行切词,通过使用停用词表和 TF-IDF 算法以及位置加权算法,将无意义的词过滤掉,提取待标引数据的特征词;根据用户标注词表,对待标引数据同时进行关键词和分类号标引,并将标引结果与人工标引结果进行对比,验证该方法的有效性。

3.2　TF-IDF 介绍

TF-IDF 算法用于评估某一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。TF 表示特征词 m 在文档 D 中出现的频率,IDF 表示所有文档中出现特征词 m 的文档数。其常用计算方法如下:

(1) $TF = \frac{m}{M}$,其中 M 表示文档 D 的总的单词数。

(2) $IDF = \log(\frac{N}{n} + 0.01)$,其中 N 为总文档数,n 为包含特征词 m 的文档数。

(3) $TF\text{-}IDF = TF \times IDF$

通过 TF-IDF 算法,能够将表示文本主要特征内容的关键词语找出来,同时将一些无意义的干扰词语过滤掉。

3.3　用户自然标注词表构建

以“万方核心期刊库”为语料库,抽取期刊论文的关键词和中图分类号,构建“用户自然标注词表”的基础库。在基础库中,需要对一个分类号对应多个关键词的情况进行处理,为每一个分类号选取出最能代表该类的关键词。

例如:分类号 1 对应的关键词有 A、B、C 三个,词频为 3、4、2;分类号 2 对应的关键词有 A、B、D 三个,词频为 2、1、2。采用 TF-IDF 算法为每个分类号选择特征词。选择过程如下:

第一步:对每个类里的关键词的词频进行归一化。

以关键词 A 为例,分类号 1 中的关键词 A 归一化后 $TF_1 = 3/9 = 0.3$,分类号 2 中的关键词 A 归一化后 $TF_2 = 2/5 = 0.4$。

第二步:计算每个类中 A 的逆分类号数。

$IDF_1 = \log(2/2 + 0.01) = 0.004$

$IDF_2 = \log(2/2 + 0.01) = 0.004$

第三步:分别计算 TF × IDF 的值。

$A_1 = TF_1 \times IDF1 = 0.001\ 2$

$A_2 = TF_2 \times IDF_2 = 0.001\ 6$

第四步:根据 TF × IDF 值的大小,确定关键词 A 对应的分类号。

由 A1 < A2,所以关键词 A 对应的分类号为分类号 2。

通过以上步骤,构建“用户自然标注词表”,并以“分类号 – 关键词”的形式存储。通过 TF-IDF 算法构建用户自然标注词表,能够将某个领域内绝大多数符合用户使用习惯的特征词选出来,实现对词表的优化。

陈白雪,宋培彦.基于用户自然标注的TF-IDF辅助标引算法及实证研究[J].图书情报工作,2018,62(1):132-139.

3.4 关键词标引和分类号标引

对待标引数据进行关键词标引和分类号标引依靠的是"用户自然标注词表"。对待标引数据进行关键词和分类号标引的主要步骤如下:

3.4.1 对待标引数据进行切词 在对待标引数据进行切词的过程中,采用的词表是"用户自然标注词表",采用IK Analyzer开源软件对待标引数据进行切词。IK Analyzer是一个开源的,基于Java语言开发的轻量级的中文分词工具包,支持用户词典扩展,能够加载"用户自然标注词表",在切词过程中,采取的是正向最大匹配算法。

3.4.2 过滤无意义的词语 在构建用户自然标注词表的过程中,同时需要构建一个停用词表。停用词表中包括一般大众通用的日常词语,不具有明显的学科或领域主题的特征,例如:"研究""作用"等一些无专指意义的词语。在切词完成后,使用停用词表将一些干扰词语排除掉,保证剩下的词语尽量有意义,能够表达待标引数据的一些内容特征。

3.4.3 关键词和分类号标引 在TF-IDF算法中,主要考虑到了词语的频次,没有考虑到词语在文本中所处的位置。因此,在关键词提取过程中,引入了位置加权法,通过对词语在文本中所处位置的不同,为不同位置的词语赋予一定的权重,体现词语对文本主题的重要程度。

通过停用词表将无意义的词语过滤掉后,对剩下的词语TF-IDF和位置加权算法,计算每个词语的得分。其计算过程如下:

(1)计算文本中所有词语的TF-IDF值,求出词语的得分。

(2)判断词语在文本中的位置,根据位置的不同,赋予一定的权重。通常情况下,词语处于关键词位置的权重较大,其次是题目,最后是摘要和正文。

(3)根据词语位置的不同,分别计算词语的TF-IDF权重值,即TF-IDF值乘以权重值。

(4)对所有词语按照TF-IDF权重值从高到低进行排序。

(5)关键词标引。为了使标引结果尽可能辅助人工标引,取得分最高的前10个词语(若不足10个,则全部保留),即为关键词标引的结果。

(6)分类号标引。将这些关键词与用户自然标注词表进行精确匹配,查找关键词对应的分类号,即可为待标引数据进行分类,获得1个推荐分类号。

3.5 辅助标引结果评测

对待标引数据同时采用以上算法和人工标引两种方法分别进行关键词标引和分类标引,从标引准确度等方面对标引结果进行对比,评测上述标引算法是否可行。

3.5.1 关键词标引结果评测 在对关键词进行对比时,引入两个统计指标,分别是:"相同比"和"相似比"。其计算方式如下:

$$\text{相同比}=\frac{\text{机标关键词与人标关键词完全相同的个数}}{\text{人标关键词}}$$ 公式(1)

$$\text{相似比}=\frac{\text{机标关键词与人标关键词互为等级或相关关系的词}+\text{机标关键词与人标关键词完全相同的个数}}{\text{人标关键词}}$$ 公式(2)

在公式1和公式2中,"机标关键词"是指通过计算机对待标引数据标引的关键词,一般为10个;"人标关键词"是指专业人员为待标引数据标引的关键词,一般为3-7个关键词。

3.5.2 分类号标引结果评测 在对分类号进行对比时,只要"机标分类号"与"人标分类号"前三位一致,即可判断"机标分类号"是合理的。例如:一条待标引数据机标的分类号是R73,人标的分类号是R737.25和R730.4,R73与R737.25和R730.4的前三位一致,因此,可将"机标分类号"与"人标分类号"视为一致,即"机标分类号"是合理的。

在该评测方法中,"机标分类号"指的是通过计算机对待标引数据标引的中图分类号;"人标分类号"指的是专业人员赋予待标引数据的中图分类号。

在以上结果评测过程中,视专业人员标引的结果是正确的。

4 实证研究

科技项目是指以科学研究和技术开发为内容而单独立项的项目。其基本的元数据字段包括项目名称、关键词、项目简介、项目负责人等字段。为了方便对科技项目数据进行统一管理,需对现有的科技项目数据进行标引、分类和整合,获取科技项目数据的关键词和分类,从而对科技项目数据进行分类和组织,而标引是对数据进行分类和组织的一个重要手段。以课题组承担的国内科研项目工作为应用场景,通过对科研项目数

据进行标引和分类，进而实现对科研项目的有效组织与服务，将会在科技项目查重与检索方面发挥有效作用。

4.1　实验过程

4.1.1　用户标注词表构建　用户标注词表的数据源选取了“万方核心期刊库”里的“U27 车辆工程”“R73 肿瘤学”“U44 桥涵工程”三个领域里的期刊论文的关键词和分类号，形成“分类号 – 关键词”列表，共计 221 664条记录。其构建过程如下：

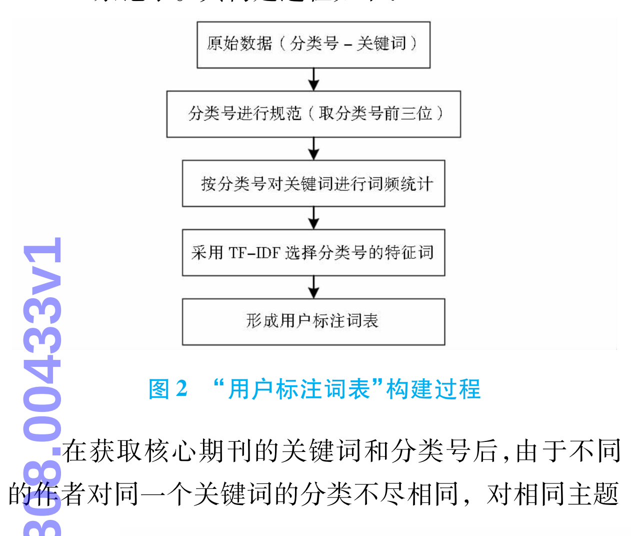


图 2　“用户标注词表”构建过程

在获取核心期刊的关键词和分类号后，由于不同的作者对同一个关键词的分类不尽相同，对相同主题的论文的分类号层级也不全相同，分类号的层级有三级、四级或五级以上，因此，在构建用户标注词表时，需要对分类号进行规范，根据期刊论文中作者赋予的分类号的位数，分类号最少是三级，可以满足大多数常规需求，因此，在构建用户标注词表时将所有分类号归到了各自的上位类“U27”“U44”和“R73”。

通过统计每个关键词在不同分类号中出现的词频，使用 TF-IDF 算法，为每个类号选出能够代表该类的关键词，最终形成 94 053 条记录，见图 3。

4.1.2　科技项目数据关键词与分类号标引　在对科技项目数据进行特征词提取时，主要通过项目标题和摘要提取。在实验过程中，随机选取三个领域的科技项目数据 840 条，先用 IK Analyzer 切词软件对项目数据进行切词，然后，使用停用词表将没有专指意义的词语过滤掉，其次，使用 TF-IDF 和位置加权算法对剩下的词语计算和排序，最后，提取科技项目数据的关键词，对科技项目数据进行关键词标引与分类号标引。根据题目和摘要对科技项目的重要性，依据经验将其权重比设为 6∶4。其部分计算结果见图 4。

	A	B	C	D	E	F
1	分类号	关键词	词频	tf	idf	source
2	U44	桥梁工程	196	0.006258181	0.4054651081081644	0.0025374740357254604
3	R73	预后	189	0.001024945	1.0986122886681098	0.0011260171722089357
4	R73	肿瘤	174	0.0009436	1.0986122886681098	0.0010366505555872284
5	R73	免疫组织化学	171	0.000927331	1.0986122886681098	0.0010187772322628869
6	R73	诊断	167	0.000905639	0	0
7	R73	细胞凋亡	160	0.000867678	1.0986122886681098	0.0009532417134069681
8	R73	凋亡	152	0.000824295	1.0986122886681098	0.0009055806164876795
9	R73	治疗	151	0.000818872	1.0986122886681098	0.0008996228420462324
10	R73	磁共振成像	137	0.00074295	1.0986122886681098	0.0008162139998659722
11	R73	免疫组化	135	0.000732104	1.0986122886681098	0.0008042984509830779
12	R73	体层摄影术，X线计算机	129	0.000699566	1.0986122886681098	0.0007685518043343949
13	R73	增殖	123	0.000667028	1.0986122886681098	0.0007328051576857119
14	R73	肿瘤转移	121	0.000656182	1.0986122886681098	0.0007208896088028176
15	R73	化疗	120	0.000650759	1.0986122886681098	0.0007149318343613705

图 3　用户标注词表截图

正题名	文摘	机标关键词	机标分类号
经动脉化疗栓塞治疗肝癌中循环内皮细胞的变化及意义	研究目标： 主要研究内容： 项目简介：	经动脉化疗栓塞 = 11.224319，循环内皮细胞 = 11.224319，肝癌 = 4.879882，	R73
BMP-2/Smad信号通路及相关因子Noggin、Smurf1与舌癌侵袭转移	研究目标： 主要研究内容： 项目简介：	舌癌 = 6.7345915，noggin = 6.7345915，smad = 6.041445，信号通路 = 4.026541，侵袭转移 = 3.4023871，	R73
SPARC、Gp60受体在乳腺癌中的表达与白蛋白结合型紫杉醇疗效关系的研究	研究目标： 主要研究内容： 项目简介：	sparc = 11.224319，白蛋白结合型紫杉醇 = 11.224319，乳腺癌 = 5.206123，	R73
经门静脉栓塞化疗（PVCE）治疗肝转移瘤的研究	研究目标： 主要研究内容： 项目简介：	肝转移瘤 = 16.83648，栓塞化疗 = 16.83648，	R73
血清miRNA作为恶性黑素瘤早期诊断标记的研究	研究目标： 主要研究内容： 项目简介：	诊断标记 = 11.224319，血清mirna = 11.224319，恶性黑素瘤 = 10.069075，	R73
热休克蛋白70功能肽-甲胎蛋白表位肽复合物抗瘤免疫机制的研究	研究目标： 主要研究内容：已证实"免疫佐剂"热休克蛋白70(HSP70)偶联抗原肽可诱发肿瘤特异性主动免疫。甲胎蛋白(AFP)作为肿瘤胚胎	表位 = 6.7345915，热休克蛋白70 = 6.7345915，甲胎蛋白 = 6.041445，免疫机制 = 6.041445，复合物 = 5.348297，hsp70 = 2.196889，afp = 1.3977692，抗原表位 = 1.2244712，免疫原性 = 1.2244712，免疫效应 = 1.0984445，	R73
西安地区女性人乳头瘤病毒感染状况与宫颈病变相关性研究	研究目标： 主要研究内容：采用基因芯片导流杂交检测法，对400例西安地区妇女宫颈癌及宫颈上皮内瘤变样本进行21种人乳头瘤病毒(human papil	人乳头瘤病毒 = 6.7345915，宫颈病变 = 6.7345915，西安地区 = 6.7345915，感染状况 = 6.7345915，相关性研究 = 4.4320064，基因芯片导流杂交 = 1.683648，宫颈癌 = 1.5358254，hpv = 1.5103612，亚型分布 = 0.841824，知晓率 = 0.841824，	R73

图 4　科技项目数据关键词与分类号标引结果部分截图

4.2 实验结果

为了验证该方法的有效性,请专业标引人员在事先不接触机标结果的前提下,人工对这 840 条数据进行关键词和分类号标引。人工标引的过程如下:

将 840 条科技项目数据的标题和文摘信息以 EXCEL 文件形式发给具有专业知识背景的标引人员;专业标引人员依据自己的背景知识,根据科技项目数据的题目和摘要,从中抽取或赋予能代表该数据内容特征的 3－7 个关键词和 1－3 个分类号。由于有些科技项目数据的摘要内容为空,专业标引人员在进行关键词和分类号标引时,只能根据题目进行标引,这样选出的关键词可能不足 3 个,这时就按照有多少标多少的原则进行标引即可。在人工标引过程中,专业人员可以主要依据《汉语主题词表》和《中国图书馆分类法》,尽可能使用较为规范的关键词和分类号对待标引数据进行标引。

人工标引的结果如图 5 所示:

正题名	文摘	人标关键词	人标分类号
经动脉化疗栓塞治疗肝癌中循环内	研究目标:　主要研究内容:　项目简	经动脉化疗栓塞治疗;肝癌;循环内皮细胞	R730.53;R735.7
BMP-2/Smad信号通路及相关因子N	研究目标:　主要研究内容:　项	BMP-2/Smad信号通路;Noggin;Smurf1;舌癌;侵袭转移	R739.86
SPARC、Gp60受体在乳腺癌中的表	研究目标:　主要研究内容:　项	SPARC受体;Gp60受体;乳腺癌;紫杉醇;临床疗效	R737.9
经门静脉栓塞化疗（PVCE）治疗肝	研究目标:　主要研究内容:　项	经门静脉栓塞化疗;肝转移瘤;治疗效果	R730.53;R735.7
血清miRNA作为恶性黑素瘤早期诊	研究目标:　主要研究内容:　项	血清miRNA;恶性黑素瘤;早期诊断	R730.4;R739.5
热休克蛋白70功能肽-甲胎蛋白表	研究目标:　主要研究内容:已证实&c	热休克蛋白70;抗瘤免疫机制;抗原肽;甲胎蛋白	R730.5
西安地区女性人乳头瘤病毒感染状	研究目标:　主要研究内容:采用基因	人乳头瘤病毒;感染状况;宫颈病变	R737.33
乙酰胆碱及其受体阻滞剂在胆管癌	研究目标:　主要研究内容:　项	乙酰胆碱;阻滞剂;胆管癌;细胞增殖;神经浸润	R735.8
雌激素及其受体ERa抑制BCG治疗	研究目标:　主要研究内容:1.应用临	雌激素;浅表性膀胱癌;抑制剂	R737.14
大肠癌转移肿瘤干细胞分子标记物	研究目标:　主要研究内容:　项	大肠癌;干细胞;分子标记物	R735.34
Tivantinib抗肝癌的靶点探索及机	研究目标:第一部分:进一步证明Tivantin	Tivantinib;抗肝癌机制;靶点	R735.7
CUTL1调控BMP/Smads通路参与恶	研究目标:　主要研究内容:转录因子	恶性黑素瘤;CUTL1调控;BMP/Smads通路;发生机制	R739.5
乳腺癌患者慢性身心应激中CACUL	研究目标:　主要研究内容:研究发现	乳腺癌患者;慢性身心应激;CACUL1信号通路;生物学作用	R737.9
miR-211在人肝细胞癌变过程中的	研究目标:　主要研究内容:肝细胞癌	miR-211;肝细胞癌;癌变过程;作用机制	R735.7

图 5　科技项目数据人工标引部分结果

由于“机标关键词”选取了每条科技项目数据的前 10 个特征词,而人工标引时为每条科技项目数据标了 3－7 个关键词,因此,在“机标关键词”与“人标关键词”进行对比时,采取了两个指标:“相同比”和“相似比”。其分析结果如图 6 所示:

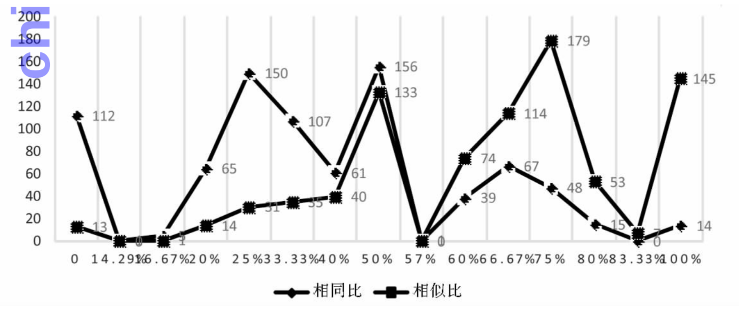


图 6　关键词标引“相同比”与“相似比”实验结果

由图 6 可知,“相同比”在 50% 以上(包括 50%)有 339 条,占总数的 40.4%,将“机标关键词”进行扩展或缩减后,即“相似比”大多都在 60% 以上,有 572 条(包括 60%),占总数的 68.1%。这样,在进行机器辅助标引时,能够将科技项目数据的关键词的相关词标引出来,再加以人工判断,即可为科技项目数据赋予符合多数用户使用习惯的关键词,标引准确度较高,且符合用户习惯。

“人标分类号”是根据专业人员的知识与背景,经过判断赋予的一个或多个分类号,而“机标分类号”是根据“用户自然标注词表”自动判断的,根据 2.3 中对分类号标引的验证方法进行检验表明,“机标分类号”与“人标分类号”前三位一致的有 705 条,占总数的 83.9%,一致性较高。其分析结果见图 7。

4.3 实验分析

4.3.1　规范分类号的效率与性能　在中文期刊核心库中,由于不同的作者对同一关键词的分类不尽相同,同一关键词可能对应多个分类号,在对科技项目数据进行分类标引时,只要能满足科技项目数据的管理需求即可,因此,在构建用户标注词表对分类号进行规范时,只取了分类号的前三位,可以基本满足科技项目的管理需求。这样在使用机器标引时,为了尽可能更客观评价标引结果,在人工标引与机器标引结果进行对比时,采用向上靠近的方法,只要“机标分类号”与“人标分类号”的前三位一致时,就认为其是正确的。虽然机器标引的分类号没有人工标引的分类号的颗粒度细,但是,机器标引的效率远远大于人工标引的效率,而且标引出的分类可以辅助人工标引。未来,在构建用户标注词表时,可以取分类号的前

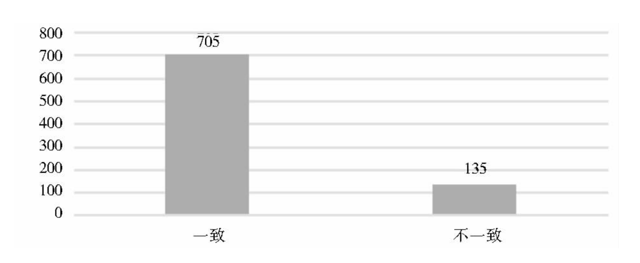


图 7 “机标分类号”与“人标分类号”前三位一致的实验结果

四位或前五位，然后再进行测试。

4.3.2 用户自然标注词表的合理性 在该算法中，“机标关键词”的正确性绝大部分取决于用户自然标注词表，在实验过程中，用户自然标注词表中的关键词主要来源于万方数据知识服务平台中核心期刊论文的关键词和分类号，最终形成 9 万多的关键词词表，但是仍不可能把三个领域中的关键词穷尽。另外，机器辅助标引属于抽词标引，而在进行人工标引时，同时采用了抽词标引和赋词标引。因此，在进行对比时，人工标引的一些关键词是无法用机器标引出来的，所以，在进行标引时，需要人机互助，用机器标引来辅助人工标引。

4.3.3 TF-IDF 的适用性 TF-IDF 广泛用于信息检索与数据挖掘方面，是一种比较成熟的算法，该算法容易理解，易于操作，在特征词选取方面具有较好的适用性。笔者通过将其运用在用户自然标注词表构建和科技项目数据特征词提取这两步，再加上停用词表，能够将一些具有干扰性的词语过滤掉，选出的特征词基本上能够描述科技项目数据的主题，同时也符合用户的标引习惯。在实验过程中，选取了 3 个学科领域进行实验，当扩大实验领域时，该算法的普适性仍需要进一步验证。

4.3.4 人工标引的主观性 在机标结果与人标结果进行对比时，考虑到人工标引的专业性，默认为人工标引的结果是正确的，实际上，不同的标引人员对于同一主题的标引会有一定的主观性，在进行关键词标引时，采用的关键词可能不尽相同，在标引颗粒度、主题倾向性方面产生偏差，例如：在肿瘤学里对于“斑马鱼模型”一词，有的采用“斑马鱼”进行标引，有的采用“动物模型”进行标引，而在机标中，采用了“斑马鱼模型”进行标引，准确性和一致性较高，可以辅助提高标引效果。因此，在进行科技项目数据辅助标引时，可以先通过自动标引，将这些关键词标出来，推荐给相关标引人员，再由标引人员进行判断，通过迭代循环，不仅可以提高机器标引的质量，也为人工标引提供了更好的辅助参考。

5 结论

笔者以中文核心期刊论文的关键词和分类号为源数据，对关键词词频进行统计，使用 TF-IDF 算法构建用户标注词表，通过 IK Analyzer 分词软件对待标引的科技项目数据进行切词，提取科技项目数据的特征词，对科技项目数据进行关键词标引和分类标引，使标引的效率和准确度有了较大提高。由实验结果可知，采用基于用户自然标注的科技项目数据机器辅助标引算法，使得“机标关键词”与“人标关键词”的相似比在 60% 以上的科技项目数据占总数的 68.1%，“机标分类号”与“人标人类号”一致的占总数的 83.9%，初步证明了该方法的有效性。通过优化 TF-IDF 计算模型、不断提高标引精度，进一步发挥用户自然标注知识库的效率，并在更多的学科领域进行验证，逐步达到应用水平，是下一步研究的重点方向。

参考文献：

[1] 马张华. 信息组织[M]. 北京：清华大学出版社，2001.
[2] 白华. 用户标注的词语网络与语义描述[J]. 图书情报工作，2010，54(2)：70－73.
[3] 孙茂松. 基于互联网自然标注资源的自然语言处理[J]. 中文信息学报，2011，25(6)：26－32.
[4] 章成志. 基于集成学习的自动标引方法研究[J]. 情报学报，2010，29(1)：3－8.
[5] 李纲，戴强斌. 基于词汇链的关键词自动标引方法[J]. 图书情报知识，2011(3)：67－71.
[6] 曹树金，周小又，陈桂鸿. 网络舆情监控系统中的主题帖自动标引及情感倾向分析研究[J]. 图书情报知识，2012(1)：66－73.
[7] 王丹，杨晓蓉. 自动标引中的歧义词消除方法研究[J]. 图书情报工作，2014，58(5)：93－97.
[8] DE CAMPOS L M，FERNÁNDEZ-LUNA J M，HUETE J F，et al. Automatic indexing from a thesaurus using Bayesian networks：application to the classification of parliamentary initiatives[C]//European conference on symbolic and quantitative approaches to reasoning and uncertainty. Berlin：Springer Berlin Heidelberg，2007：865－877.
[9] MEDELYAN O，WITTEN I H. Thesaurus based automatic keyphrase indexing[C]//Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries. New York：ACM，2006：296－297.
[10] MERROUNI Z A，FRIKH B，OUHBI B. Automatic keyphrase extraction：an overview of the state of the art[C]// IEEE international colloquium on information science and technology. Piscataway：

IEEE, 2017:306－313.

[11] 李枫林,张景. 基于用户标注行为的相关性分析及重排序[J]. 情报理论与实践, 2010(10):57－61.

[12] 吴丹,许小梅. 图书馆与图书分享网站的用户标注行为比较研究[J]. 图书情报知识, 2013(1):85－93.

[13] 谢佳琳,张晋朝. 高校图书馆用户标注行为研究——以信息系统成功模型为视角[J]. 图书馆论坛, 2014(11):87－93.

[14] PATTERSON J, DOUGALL S, MOODY N. Systems and methods for manipulating user annotations in electronic books: United States Patent,8520025[P]. 2013－08－27.

[15] ZARRO M A, ALLEN R B. User-contributed annotations for libraries and cultural institutions[EB/OL]. [2017－06－26]. http://mikezarro.com/docs/Zarro-LRS-V-Poster.pdf.

[16] ZHANG Y Y, ZHANG C Z, CHEN G, et al. Analyzing scientific user tagging behavior on academic blogs according to tag's content characteristics - a preliminary study[EB/OL]. [2017－06－26]. https://www.ideals.illinois.edu/bitstream/handle/2142/96741/3.62_419_Zhang-Analyzing%20scientific%20user%20tagging%20behavior%20on%20academic%20blogs%20according.pdf?sequence=1&isAllowed=y.

[17] PAN X, HE S, ZHU X, et al. How users employ various popular tags to annotate resources in social tagging: an empirical study[J]. Journal of the Association for Information Science & Technology, 2016, 67(5):1121－1137.

[18] 马费成,张斌. 图书标注环境下用户的认知特征[J]. 中国图书馆学报, 2014(1):4－14.

[19] 常唯. 论网络环境下用户标注的价值与应用[J]. 图书情报工作, 2008, 52(1):9－12.

[20] AIZAWA A. An information-theoretic perspective of tf-idf measures[J]. Information processing and management, 2003, 39(1):45－65.

[21] 路永和, 李焰锋. 改进 TE-IDF 算法的文本特征项权值计算方法[J]. 图书情报工作, 2013, 57(3):90－95.

[22] 覃世安, 李法运. 文本分类中 TF-IDF 方法的改进研究[J]. 现代图书情报技术, 2013, 29(10):27－30.

[23] 刘勘,周丽红,陈譞. 基于关键词的科技文献聚类研究[J]. 图书情报工作, 2012, 56(4):6－11.

[24] SAMOYLOV A B. Evaluation of the delta TF-IDF features for sentiment analysis[C]//International conference on analysis of images, social networks and texts_x000D_. Berlin: Springer, 2014: 207－212.

[25] PHILIP S, SHOLA P B, OVYE A. Application of content-based approach in research paper recommendation system for a digital library[J]. International journal of advanced computer science & applications, 2014, 5(10):37－40.

[26] Xu R. POS weighted TF-IDF algorithm and its application for an MOOC search engine[C]// International conference on audio, language and image processing. Piscataway: IEEE, 2015:868－873.

[27] DADGAR S M H, ARAGHI M S, FARAHANI M M. A novel text mining approach based on TF-IDF and support vector machine for news classification[EB/OL]. [2017－06－26]. http://ieeexplore.ieee.org/stamp/stamp.jsp? arnumber=7569223.

作者贡献说明：

陈白雪：论文框架设计、数据实验、初稿撰写；

宋培彦：实验结果分析、论文修改及定稿。

Empirical Research on TF-IDF Assisted Indexing Algorithm Based on Users' Natural Annotation

Chen Baixue　Song Peiyan

Institute of Scientific and Technical Information of China, Beijing 100038

Abstract: [**Purpose/significance**] This paper studies the TF-IDF assisted indexing algorithm based on the user natural annotation from the users' point of view. [**Method/process**] First, the keywords and the classification number in Chinese core journals were taken as the data source. The user natural annotation vocabulary was constructed by computing the keywords frequency and using the TF-IDF algorithm. Second, the featured words were extracted from the scientific and technological project data by the IK Analyzer word segmentation software and the TF-IDF algorithm. Finally, the keywords and classification number of the scientific and technological project data were indexed synchronously. [**Result/conclusion**] The experiment indicates that the data of scientific and technical projects take up 68.1% in total. In these projects, the ratio similitude of the keywords of machine indexing and the keywords of human indexing is more than 60% in total. The ratio of the uniformity in the former three numbers of machine-indexed classification number and the human-indexed classification number is 83.9% in total. It is feasible to adopt the TF-IDF algorithm based on the users' natural annotation data.

Keywords: assisted indexing　user natural annotation　TF-IDF algorithm　information organization